

PC ALGORITHM FOR GAUSSIAN COPULA GRAPHICAL MODELS

NAFTALI HARRIS AND MATHIAS DRTON

ABSTRACT. The PC algorithm uses conditional independence tests for model selection in graphical modeling with acyclic directed graphs. In Gaussian models, tests of conditional independence are typically based on Pearson correlations, and high-dimensional consistency results have been obtained for the PC algorithm in this setting. We prove that high-dimensional consistency carries over to the broader class of Gaussian copula or *nonparanormal* models when using rank-based measures of correlation. For graphs with bounded degree, our result is as strong as prior Gaussian results. In simulations, the ‘Rank PC’ algorithm works as well as the ‘Pearson PC’ algorithm for normal data and considerably better for non-normal Gaussian copula data, all the while incurring a negligible increase of computation time. Simulations with contaminated data show that rank correlations can also perform better than other robust estimates considered in previous work when the underlying distribution does not belong to the nonparanormal family.

1. INTRODUCTION

Let $G = (V, E)$ be an acyclic digraph with finite vertex set. We will typically write $v \rightarrow w \in E$ to indicate that (v, w) is an edge in E . The digraph G determines a statistical model for the joint distribution of a random vector $X = (X_v)_{v \in V}$ by requiring that X satisfy conditional independence relations that are natural if the edges in E encode causal relationships among the random variables X_v . We refer the reader to [Lau96, Pea09, SGS00] or [DSS09, Chap. 3] for background on statistical modeling with directed graphs. As common in this field, we use the abbreviation DAG (for ‘directed acyclic graph’) to refer to acyclic digraphs.

The conditional independences associated with the graph G may be determined using the concept of d-separation. Since a DAG contains at most one edge between any two nodes, we may define a path from a node u to a node v to be a sequence of distinct nodes (v_0, v_1, \dots, v_n) such that $v_0 = u$, $v_n = v$ and for all $1 \leq k \leq n$, either $v_{k-1} \rightarrow v_k \in E$ or $v_{k-1} \leftarrow v_k \in E$. Two distinct nodes u and v are then said to be *d-separated* by a set $S \subset V \setminus \{v, u\}$ if every path from u to v contains three consecutive nodes (v_{k-1}, v_k, v_{k+1}) for which one of the following is true:

- (i) The three nodes form a chain $v_{k-1} \rightarrow v_k \rightarrow v_{k+1}$, a chain $v_{k-1} \leftarrow v_k \leftarrow v_{k+1}$, or a fork $v_{k-1} \leftarrow v_k \rightarrow v_{k+1}$, and the middle node v_k is in S .
- (ii) The three nodes form a collider $v_{k-1} \rightarrow v_k \leftarrow v_{k+1}$, and neither v_k nor any of its descendants is in S .

Suppose A, B, S are pairwise disjoint subsets of V . Then S d-separates A and B if S d-separates any pair of nodes a and b with $a \in A$ and $b \in B$. Finally, the

Key words and phrases. Copula, covariance matrix, graphical model, model selection, multivariate normal distribution, nonparanormal distribution.

joint distribution of the random vector $X = (X_v)_{v \in V}$ is *Markov* with respect to a DAG G if X_A and X_B are conditionally independent given X_S for any triple of pairwise disjoint subsets $A, B, S \subset V$ such that S d-separates A and B in G . Here, X_A denotes the subvector $(X_v)_{v \in A}$. It is customary to denote conditional independence of X_A and X_B given X_S by $X_A \perp\!\!\!\perp X_B \mid X_S$.

We will be concerned with the consistency of an algorithm for inferring a DAG from data. Graph inference is complicated by the fact that two DAGs $G = (V, E)$ and $H = (V, F)$ with the same vertex set V may be *Markov equivalent*, that is, they may possess the same d-separation relations and, consequently, induce the same statistical model. To give an example, the graphs $u \rightarrow v \rightarrow w$ and $u \leftarrow v \leftarrow w$ are Markov equivalent, but $u \rightarrow v \rightarrow w$ and $u \rightarrow v \leftarrow w$ are not. As first shown in [VP91], two DAGs G and H are Markov equivalent if and only if they have the same skeleton and the same unshielded colliders. The *skeleton* of a digraph G is the undirected graph obtained by converting each directed edge into an undirected edge. An *unshielded collider* is a triple of nodes (u, v, w) that induces the subgraph $u \rightarrow v \leftarrow w$, that is, there is no edge between u and w .

Let $[G]$ be the Markov equivalence class of an acyclic digraph $G = (V, E)$. Write $E(H)$ for the edge set of a DAG H , and define the edge set

$$[E] = \bigcup_{H \in [G]} E(H).$$

That is, $(v, w) \in [E]$ if there exists a DAG $H \in [G]$ with the edge $v \rightarrow w$ in its edge set. We interpret the presence of both (v, w) and (w, v) in $[E]$ as an undirected edge between v and w . Following the most closely related literature, we call the graph $C(G) = (V, [E])$ the *completed partially directed acyclic graph* (CPDAG) for G , but other terminology such as the *essential graph* is in use. The graph $C(G)$ is partially directed as it may contain both directed and undirected edges, and it is acyclic in the sense of its directed subgraph having no directed cycles. Two DAGs G and H satisfy $C(G) = C(H)$ if and only if $[G] = [H]$, making the CPDAG a useful graphical representation of a Markov equivalence class; see [AMP97, Chi02].

The PC algorithm, named for its inventors Peter Spirtes and Clark Glymour, uses conditional independence tests to infer a CPDAG from data [SGS00]. In its population version, the algorithm amounts to a clever scheme to reconstruct the CPDAG $C(G)$ from answers to queries about d-separation relations in the underlying DAG G . Theorem 1 summarizes the properties of the PC algorithm that are relevant for the present paper. For a proof of the theorem as well as a compact description of the PC algorithm we refer the reader to [KB07]. Recall that the degree of a node is the number of edges it is incident to, and that the degree of a DAG G is the maximum degree of any node, which we denote by $\deg(G)$.

Theorem 1. *Given only the ability to check d-separation relations in a DAG G , the PC algorithm finds the CPDAG $C(G)$ by checking whether pairs of distinct nodes are d-separated by sets S of cardinality $|S| \leq \deg(G)$.*

The joint distribution of a random vector $X = (X_v)_{v \in V}$ is *faithful* to the DAG G if, for any triple of pairwise disjoint subsets $A, B, S \subset V$, we have that S d-separates A and B in G if and only if $X_A \perp\!\!\!\perp X_B \mid X_S$. Under faithfulness, statistical tests of conditional independence can be used to determine d-separation relations in a DAG and lead to a sample version of the PC algorithm that is applicable to data.

If X follows the multivariate normal distribution $N(\mu, \Sigma)$, with positive definite covariance matrix Σ , then

$$(1.1) \quad X_A \perp\!\!\!\perp X_B \mid X_S \iff X_u \perp\!\!\!\perp X_v \mid X_S \quad \forall u \in A, v \in B.$$

Moreover, the pairwise conditional independence of X_u and X_v given X_S is equivalent to the vanishing of the *partial correlation* $\rho_{uv|S}$, that is, the correlation obtained from the bivariate normal conditional distribution of (X_u, X_v) given X_S . The iterations of the PC algorithm make use of the recursion

$$(1.2) \quad \rho_{uv|S} = \frac{\rho_{uv|S \setminus w} - \rho_{uw|S \setminus w} \rho_{vw|S \setminus w}}{\sqrt{(1 - \rho_{uw|S \setminus w}^2)(1 - \rho_{vw|S \setminus w}^2)}},$$

where $w \in S$, and we define $\rho_{uv|\emptyset} = \rho_{uv}$ to be correlation of u and v . Our later theoretical analysis will use the fact that

$$(1.3) \quad \rho_{uv|S} = -\frac{\Psi_{uv}^{-1}}{\sqrt{\Psi_{uu}^{-1} \Psi_{vv}^{-1}}},$$

where $\Psi = \Sigma_{(u,v,S),(u,v,S)}$ is the concerned principal submatrix of Σ . A natural estimate of $\rho_{uv|S}$ is the sample partial correlation obtained by replacing Σ with the empirical covariance matrix of available observations. Sample partial correlations derived from independent normal observations have favorable distributional properties [And03, Chap. 4], which form the basis for the work of [KB07] who treat the PC algorithm in the Gaussian context with conditional independence tests based on sample partial correlations. The main results in [KB07] show high-dimensional consistency of the PC algorithm, when the observations form a sample of independent normal random vectors that are faithful to a suitably sparse DAG.

The purpose of this paper is to show that the PC algorithm has high-dimensional consistency properties for a broader class of distributions, when standard Pearson-type empirical correlations are replaced by rank-based measures of correlations in tests of conditional independence. The broader class we consider comprises the distributions with Gaussian copula. Phrased in the terminology of [LLW09], we consider *nonparanormal* distributions. Recall that a correlation matrix is a covariance matrix with all diagonal entries equal to one.

Definition 1. Let $f = (f_v)_{v \in V}$ be a collection of strictly increasing, but not necessarily continuous functions $f_v : \mathbb{R} \rightarrow \mathbb{R}$, and let $\Sigma \in \mathbb{R}^{V \times V}$ be a positive definite correlation matrix. The nonparanormal distribution $NPN(f, \Sigma)$ is the distribution of the random vector $(f_v(Z_v))_{v \in V}$ for $(Z_v)_{v \in V} \sim N(0, \Sigma)$.

Taking the functions f_v to be affine shows that all multivariate normal distributions are also nonparanormal. If $X \sim NPN(f, \Sigma)$, then the univariate marginal distribution for a coordinate, say X_v , may have any continuous cumulative distribution function F , as we may take $f_v = F^{-1} \circ \Phi$, where Φ is the standard normal distribution function.

Definition 2. The Gaussian copula graphical model $NPN(G)$ associated with a DAG G is the set of all distributions $NPN(f, \Sigma)$ that are Markov with respect to G .

Since the marginal transformations f_v are deterministic, the dependence structure in a nonparanormal distribution corresponds to that in the underlying latent multivariate normal distribution. In other words, if $X \sim NPN(f, \Sigma)$ and

$Z \sim N(0, \Sigma)$, then it holds for any triple of pairwise disjoint sets $A, B, S \subset V$ that

$$(1.4) \quad X_A \perp\!\!\!\perp X_B \mid X_S \iff Z_A \perp\!\!\!\perp Z_B \mid Z_S.$$

Hence, for two nodes u and v and a separating set $S \subset V \setminus \{u, v\}$, it holds that

$$(1.5) \quad X_u \perp\!\!\!\perp X_v \mid X_S \iff \rho_{uv|S} = 0,$$

with $\rho_{uv|S}$ calculated from Σ as in (1.2) or (1.3). In light of this equivalence, we will occasionally speak of a correlation matrix Σ being Markov or faithful to a DAG, meaning that the requirement holds for any distribution $NPN(f, \Sigma)$.

In the remainder of the paper we study the PC algorithm in the nonparanormal context, proposing the use of Spearman's rank correlation and Kendall's τ for estimation of the correlation matrix parameter of a nonparanormal distribution. In Section 2, we review how transformations of Spearman's rank correlation and Kendall's τ yield accurate estimators of the latent Gaussian correlations. In particular, we summarize tail bounds from [LHY⁺12]. Theorem 2 in Section 3 gives our main result, an error bound for the output of the PC algorithm when correlations are used to determine nonparanormal conditional independence. In Corollary 1, we describe high-dimensional asymptotic scenarios and suitable conditions that lead to consistency of the PC algorithm. The proof of Theorem 2 is given in Section 4. Our simulations in Section 5 make a strong case for the use of rank correlations in the PC algorithm. Some concluding remarks are given in Section 6.

2. RANK CORRELATIONS

Let (X, Y) be a pair of random variables, and let F and G be the cumulative distribution functions of X and Y , respectively. Spearman's ρ for the bivariate distribution of (X, Y) is defined as

$$(2.1) \quad \rho^S = \text{Corr}(F(X), G(Y)),$$

that is, it is the ordinary Pearson correlation between the quantiles $F(X)$ and $G(Y)$. Another classical measure of correlation is Kendall's τ , defined as

$$(2.2) \quad \tau = \text{Corr}(\text{sign}(X - X'), \text{sign}(Y - Y'))$$

where (X', Y') is an independent copy of (X, Y) .

Suppose $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent pairs of random variables, each pair distributed as (X, Y) . Let $\text{rank}(X_i)$ be the rank of X_i among X_1, \dots, X_n . In the nonparanormal setting, the marginal distributions are continuous so that ties occur with probability zero, making ranks well-defined. The natural estimator of ρ^S is the sample correlation among ranks, that is,

$$(2.3) \quad \hat{\rho}^S = \frac{\frac{1}{n} \sum_{i=1}^n \left(\frac{\text{rank}(X_i)}{n+1} - \frac{1}{2} \right) \left(\frac{\text{rank}(Y_i)}{n+1} - \frac{1}{2} \right)}{\sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{\text{rank}(X_i)}{n+1} - \frac{1}{2} \right)^2} \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{\text{rank}(Y_i)}{n+1} - \frac{1}{2} \right)^2}}$$

$$(2.4) \quad = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (\text{rank}(X_i) - \text{rank}(Y_i))^2,$$

which can be computed in $O(n \log n)$ time. Kendall's τ may be estimated by

$$(2.5) \quad \hat{\tau} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \text{sign}(X_i - X_j) \text{sign}(Y_i - Y_j).$$

A clever algorithm using sorting and binary trees to compute $\hat{\tau}$ in time $O(n \log n)$ instead of the naive $O(n^2)$ time has been developed by [Chr05].

It turns out that simple trigonometric transformations of $\hat{\rho}^S$ and $\hat{\tau}$ are excellent estimators of the population Pearson correlation for multivariate normal data. In particular, [LHY⁺12] show that if (X, Y) are bivariate normal with $\text{Corr}(X, Y) = \rho$, then

$$(2.6) \quad \mathbb{P} \left(\left| 2 \sin \left(\frac{\pi}{6} \hat{\rho}^S \right) - \rho \right| > \epsilon \right) \leq 2 \exp \left(-\frac{2}{9\pi^2} n \epsilon^2 \right)$$

and

$$(2.7) \quad \mathbb{P} \left(\left| \sin \left(\frac{\pi}{2} \hat{\tau} \right) - \rho \right| > \epsilon \right) \leq 2 \exp \left(-\frac{2}{\pi^2} n \epsilon^2 \right).$$

Clearly, $\hat{\rho}^S$ and $\hat{\tau}^K$ depend on the observations $(X_1, Y_1), \dots, (X_n, Y_n)$ only through their ranks. Since ranks are preserved under strictly increasing functions, (2.6) and (2.7) still hold if $(X, Y) \sim \text{NPN}(f, \Sigma)$ with Pearson correlation $\rho = \Sigma_{xy}$ in the underlying latent bivariate normal distribution. Throughout the rest of this paper, we will assume that we have some estimator $\hat{\rho}$ of ρ which has the property that, for nonparanormal data,

$$(2.8) \quad \mathbb{P}(|\hat{\rho} - \rho| > \epsilon) < A \exp(-Bn\epsilon^2)$$

for fixed constants $0 < A, B < \infty$. As just argued, the estimators considered in (2.6) and (2.7) both have this property.

When presented with multivariate observations from a distribution $\text{NPN}(f, \Sigma)$, we apply the estimator from (2.8) to every pair of coordinates to obtain an estimator $\hat{\Sigma}$ of the correlation matrix parameter. Plugging $\hat{\Sigma}$ into (1.2) or equivalently into (1.3) gives partial correlation estimators that we denote $\hat{\rho}_{uv|S}$.

3. RANK PC ALGORITHM

Based on the equivalence (1.5), we may use the rank-based partial correlation estimates $\hat{\rho}_{uv|S}$ to test conditional independences. In other words, we conclude that

$$(3.1) \quad X_u \perp\!\!\!\perp X_v | X_S \iff |\hat{\rho}_{uv|S}| \leq \gamma,$$

where $\gamma \in [0, 1]$ is a fixed threshold. We will refer to the PC algorithm that uses the conditional independence tests from (3.1) as the ‘Rank PC’ (RPC) algorithm. We write $\hat{C}_\gamma(G)$ for the output of the RPC algorithm with tuning parameter γ .

The RPC algorithm consist of two parts. The first part computes the correlation matrix $\hat{\Sigma} = (\hat{\rho}_{uv})$ in time $O(p^2 n \log n)$, where $p := |V|$. This computation takes $O(\log n)$ longer than its analogue under use of Pearson correlations. The second part of the algorithm is independent of the type of correlations involved. It determines partial correlations and performs graphical operations. For an accurate enough estimate of a correlation matrix Σ that is faithful to a DAG G , this second part takes $O(p^{\deg(G)})$ time in the worst case, but it is often much faster; compare [KB07]. For high-dimensional data with n smaller than p , the computation time for RPC is dominated by the second part, the PC-algorithm component. Moreover, in practice, one may wish to apply RPC for several different values of γ , in which case the estimate $\hat{\Sigma}$ needs to be calculated only once. As a result, Rank PC takes only marginally longer to compute than Pearson PC for high-dimensional data.

What follows is our main result about the correctness of RPC, which we prove in Section 4. For a correlation matrix $\Sigma \in \mathbb{R}^{V \times V}$, let

$$(3.2) \quad c_{\min}(\Sigma) := \min \{ |\rho_{uv|S}| : u, v \in V, S \subseteq V \setminus \{u, v\}, \rho_{uv|S} \neq 0 \}$$

be the minimal magnitude of any non-zero partial correlation, and let $\lambda_{\min}(\Sigma)$ be the minimal eigenvalue. Then for any integer $q \geq 2$, let

$$(3.3) \quad c_{\min}(\Sigma, q) := \min \{ c_{\min}(\Sigma_{I,I}) : I \subseteq V, |I| = q \}, \quad \text{and}$$

$$(3.4) \quad \lambda_{\min}(\Sigma, q) := \min \{ \lambda_{\min}(\Sigma_{I,I}) : I \subseteq V, |I| = q \}$$

be the minimal magnitude of a non-zero partial correlation and, respectively, the minimal eigenvalue of any $q \times q$ principal submatrix of Σ . Note that if $I \subset J$ then $c_{\min}(\Sigma_{I,I}) \leq c_{\min}(\Sigma_{J,J})$ and $\lambda_{\min}(\Sigma_{I,I}) \leq \lambda_{\min}(\Sigma_{J,J})$.

Theorem 2 (Error bound for RPC-algorithm). *Let X_1, \dots, X_n be a sample of independent observations drawn from a nonparanormal distribution $NPN(f, \Sigma)$ that is faithful to a DAG G with p nodes. For $q := \deg(G) + 2$, let $c := c_{\min}(\Sigma, q)$ and $\lambda := \lambda_{\min}(\Sigma, q)$. If $n > q$, then there exists a threshold $\gamma \in [0, 1]$ for which*

$$\mathbb{P}(\hat{C}_{\gamma}(G) \neq C(G)) \leq \frac{A}{2} p^2 \exp\left(-\frac{B\lambda^4 n c^2}{36q^2}\right),$$

where $0 < A, B < \infty$ are the constants from (2.8).

We remark that while all subsets of size q appear in the definitions in (3.3) and (3.4), our proof of Theorem 2 only requires the corresponding minima over those principal submatrices that are actually inverted in the run of the PC-algorithm.

From the probability bound in Theorem 2, we may deduce high-dimensional consistency of RPC. For two positive sequences (s_n) and (t_n) , we write $s_n = O(t_n)$ if $s_n \leq Mt_n$, and $s_n = \Omega(t_n)$ if $s_n \geq Mt_n$ for a constant $0 < M < \infty$.

Corollary 1 (Consistency of RPC-algorithm). *Let (G_n) be a sequence of DAGs. Let p_n be the number of nodes of G_n , and let $q_n = \deg(G_n) + 2$. Suppose (Σ_n) is a sequence of $p_n \times p_n$ correlation matrices, with Σ_n faithful to G_n . Suppose further that there are constants $0 \leq a, b, d, f < 1$ that govern the growth of the graphs as*

$$\log p_n = O(n^a), \quad q_n = O(n^b),$$

and minimal signal strengths and eigenvalues as

$$c_{\min}(\Sigma_n, q_n) = \Omega(n^{-d}), \quad \lambda_{\min}(\Sigma_n, q_n) = \Omega(n^{-f}).$$

If $a + 2b + 2d + 4f < 1$, then there exists a sequence of thresholds γ_n for which

$$\lim_{n \rightarrow \infty} \mathbb{P}(\hat{C}_{\gamma_n}(G_n) = C(G_n)) = 1,$$

where $\hat{C}_{\gamma_n}(G_n)$ is the output of the RPC algorithm for a sample of independent observations X_1, \dots, X_n from a nonparanormal distribution $NPN(\cdot, \Sigma_n)$.

Proof. By Theorem 2, for large enough n , we can pick a threshold γ_n such that

$$(3.5) \quad \mathbb{P}(\hat{C}_{\gamma_n}(G_n) \neq C(G_n)) \leq A' \exp(2n^a - B'n^{1-2b-2d-4f})$$

for constants $0 < A', B' < \infty$. The bound goes to zero if $1 - 2b - 2d - 4f > a$. \square

As previously mentioned, [KB07] prove a similar consistency result in the Gaussian case. Whereas our proof consists of propagation of errors from correlation to partial correlation estimates, their proof appeals to Fisher's result that under Gaussianity, sample partial correlations follow the same type of distribution as sample correlations when the sample size is adjusted by subtracting the cardinality of the conditioning set [And03, Chap. 4]. It is then natural to work with a bound on the partial correlations associated with small conditioning sets. More precisely, [KB07] assume that there is a constant $0 \leq M < 1$ such that for any n , the partial correlations $\rho_{uv|S}$ of the matrix Σ_n satisfy

$$(3.6) \quad |\rho_{uv|S}| \leq M \quad \forall u, v \in V, S \subseteq V \setminus \{u, v\}, |S| \leq q_n.$$

It is then no longer necessary to involve the minimal eigenvalues from (3.4). The work in [KB07] is thus free of an analogue to our constant f . Stated for the case of polynomial growth of p_n (with $a = 0$), their result gives consistency when $b + 2d < 1$, whereas our condition requires $2b + 2d < 1$ even if $f = 0$. (Note that our constant b corresponds to $1 - b$ in [KB07].)

In the important special case of bounded degree, however, our nonparanormal result is just as strong as the previously established Gaussian consistency guarantee. Staying with polynomial growth of p_n , i.e., $a = 0$, suppose the sequence of graph degrees $\deg(G_n)$ is indeed bounded by a fixed constant, say $q_0 - 2$. Then clearly, $b = 0$. Moreover, the set of correlation matrices of size q_0 satisfying (3.6) with $q_n = q_0$ is compact. Since the smallest eigenvalue is a continuous function, the infimum of all eigenvalues of such matrices is achieved for some invertible matrix. Hence, the smallest eigenvalue is bounded away from zero, and we conclude that $f = 0$. Corollary 1 thus implies consistency if $2d < 1$, or if $d < \frac{1}{2} = \frac{1-b}{2}$, precisely as in [KB07]. (No generality is lost by assuming $a = 0$; in either one of the compared results this constant is involved solely in a union bound over order p^2 events.)

4. PROOF OF THE ERROR BOUND

In this section, we prove the error bound in Theorem 2. Our argument starts from a uniform bound on the error in our estimate $\hat{\Sigma}$. Then we analyze how this error propagates to the partial correlation estimates $\hat{\rho}_{uv|S}$, giving again a uniform error bound. We begin by proving three lemmas about the error propagation.

The first lemma invokes classical results on error propagation in matrix inversion. Let $\|A\|$ denote the spectral norm of a matrix $A = (a_{ij}) \in \mathbb{R}^{q \times q}$, that is, $\|A\|^2$ is the maximal eigenvalue of $A^T A$. Write the l_∞ vector norm of A as

$$\|A\|_\infty = \max_{1 \leq i, j \leq q} |a_{ij}|.$$

Lemma 1 (Errors in matrix inversion). *Suppose $\Sigma \in \mathbb{R}^{q \times q}$ is an invertible matrix with minimal eigenvalue λ_{\min} . If $E \in \mathbb{R}^{q \times q}$ is a matrix of errors with $\|E\|_\infty < \epsilon < \lambda_{\min}/q$, then $\Sigma + E$ is invertible and*

$$\|(\Sigma + E)^{-1} - \Sigma^{-1}\|_\infty \leq \frac{q\epsilon/\lambda_{\min}^2}{1 - q\epsilon/\lambda_{\min}}.$$

Proof. First, note that

$$(4.1) \quad \|E\|_\infty \leq \|E\| \leq q\|E\|_\infty;$$

see entries (2, 6) and (6, 2) in the table on p. 314 in [HJ90]. Using the submultiplicativity of a matrix norm, the second inequality in (4.1), and our assumption on ϵ , we find that

$$(4.2) \quad \|E\Sigma^{-1}\| \leq \|\Sigma^{-1}\| \cdot \|E\| < \frac{q\epsilon}{\lambda_{\min}} < 1.$$

As discussed in [HJ90, Sect. 5.8], this implies that $I + E\Sigma^{-1}$ and thus also $\Sigma + E$ is invertible. Moreover, by the first inequality in (4.1) and inequality (5.8.2) in [HJ90], we obtain that

$$(4.3) \quad \|(\Sigma + E)^{-1} - \Sigma^{-1}\|_{\infty} \leq \|(\Sigma + E)^{-1} - \Sigma^{-1}\| \leq \|\Sigma^{-1}\| \cdot \frac{\|E\Sigma^{-1}\|}{1 - \|E\Sigma^{-1}\|}.$$

Since the function $x \mapsto x/(1-x)$ is increasing for $x < 1$, our claim follows from the fact that $\|\Sigma^{-1}\| = 1/\lambda_{\min}$ and the inequality $\|E\Sigma^{-1}\| < q\epsilon/\lambda_{\min}$ from (4.2). \square

Lemma 2 (Diagonal of inverted correlation matrix). *If $\Sigma \in \mathbb{R}^{q \times q}$ is a positive definite correlation matrix, then the diagonal entries of $\Sigma^{-1} = (\sigma^{ij})$ satisfy $\sigma^{ii} \geq 1$.*

Proof. The claim is trivial for $q = 1$. So assume $q \geq 2$. By symmetry, it suffices to consider the entry σ^{qq} , and we partition the matrix as

$$(4.4) \quad \Sigma = \begin{pmatrix} A & b \\ b^T & 1 \end{pmatrix}$$

with $A \in \mathbb{R}^{(q-1) \times (q-1)}$ and $b \in \mathbb{R}^{q-1}$. By the Schur complement formula for the inverse of a partitioned matrix,

$$\sigma^{qq} = \frac{1}{1 - b^T A^{-1} b};$$

compare [HJ90, §0.7.3]. Since A is positive definite, so is A^{-1} . Hence, $b^T A^{-1} b \geq 0$. Since Σ^{-1} is positive definite, σ^{qq} cannot be negative, and so we deduce that $\sigma^{qq} \geq 1$, with equality if and only if $b = 0$. \square

The next lemma treats the error propagation from the inverse of a correlation matrix to partial correlations.

Lemma 3 (Error in partial correlations). *Let $A = (a_{ij})$ and $B = (b_{ij})$ be symmetric 2×2 matrices. If A is positive definite with $a_{11}, a_{22} \geq 1$ and $\|A - B\|_{\infty} < \delta < 1$, then*

$$\left| \frac{a_{12}}{\sqrt{a_{11}a_{22}}} - \frac{b_{12}}{\sqrt{b_{11}b_{22}}} \right| < \frac{2\delta}{1 - \delta}.$$

Proof. Without loss of generality, suppose $a_{12} \geq 0$. Since $\|A - B\|_{\infty} < \delta$,

$$\begin{aligned} \frac{b_{12}}{\sqrt{b_{11}b_{22}}} - \frac{a_{12}}{\sqrt{a_{11}a_{22}}} &< \frac{a_{12} + \delta}{\sqrt{(a_{11} - \delta)(a_{22} - \delta)}} - \frac{a_{12}}{\sqrt{a_{11}a_{22}}} \\ &= \frac{\delta}{\sqrt{(a_{11} - \delta)(a_{22} - \delta)}} + a_{12} \left(\frac{1}{\sqrt{(a_{11} - \delta)(a_{22} - \delta)}} - \frac{1}{\sqrt{a_{11}a_{22}}} \right). \end{aligned}$$

Using that $a_{11}, a_{22} \geq 1$ to bound the first term and $a_{12}^2 < a_{11}a_{22}$ to bound the second term, we obtain that

$$\begin{aligned} \frac{b_{12}}{\sqrt{b_{11}b_{22}}} - \frac{a_{12}}{\sqrt{a_{11}a_{22}}} &< \frac{\delta}{1-\delta} + \sqrt{a_{11}a_{22}} \left(\frac{1}{\sqrt{(a_{11}-\delta)(a_{22}-\delta)}} - \frac{1}{\sqrt{a_{11}a_{22}}} \right) \\ &= \frac{\delta}{1-\delta} + \left(\sqrt{\frac{a_{11}}{a_{11}-\delta} \cdot \frac{a_{22}}{a_{22}-\delta}} - 1 \right). \end{aligned}$$

Since the function $x \mapsto x/(x-\delta)$ is decreasing, we may use our assumption that $a_{11}, a_{22} \geq 1$ to get the bound

$$\frac{b_{12}}{\sqrt{b_{11}b_{22}}} - \frac{a_{12}}{\sqrt{a_{11}a_{22}}} < \frac{\delta}{1-\delta} + \left(\sqrt{\frac{1}{1-\delta} \cdot \frac{1}{1-\delta}} - 1 \right) = \frac{2\delta}{1-\delta}$$

A similar argument yields that

$$(4.5) \quad \frac{a_{12}}{\sqrt{a_{11}a_{22}}} - \frac{b_{12}}{\sqrt{b_{11}b_{22}}} < \frac{2\delta}{1+\delta},$$

from which our claim follows. \square

We are now ready to prove our main result.

Proof of Theorem 2. We will show that our claimed probability bound for the event $\hat{C}_\gamma(G) \neq C(G)$ holds when the threshold in the RPC algorithm is $\gamma = c/2$. By Theorem 1, if all conditional independence tests for conditioning sets of size $|S| \leq \deg(G)$ make correct decisions, then the output of the RPC algorithm $\hat{C}_\gamma(G)$ is equal to the CPDAG $C(G)$. When $\gamma = c/2$, the conditional independence test accepts a hypothesis $X_u \perp\!\!\!\perp X_v | X_S$ if and only if $|\hat{\rho}_{uv|S}| < \gamma = c/2$. Hence, the test makes a correct decision if $|\hat{\rho}_{uv|S} - \rho_{uv|S}| < c/2$ because all non-zero partial correlations for $|S| \leq \deg(G)$ are bounded away from zero by c ; recall (3.2) and (3.3). It remains to show, using the error analysis from Lemmas 1 and 3, that the event $|\hat{\rho}_{uv|S} - \rho_{uv|S}| \geq c/2$ occurs with small enough probability when $|S| \leq \deg(G)$.

Suppose our correlation matrix estimate $\hat{\Sigma} = (\hat{\rho}_{uv})$ satisfies $\|\hat{\Sigma} - \Sigma\|_\infty < \epsilon$ for

$$(4.6) \quad \epsilon = \frac{c\lambda^2}{(4+c)q + \lambda cq} > 0.$$

Choose any two nodes $u, v \in V$ and a set $S \subseteq V \setminus \{u, v\}$ with $|S| \leq \deg(G) = q-2$. Let $I = \{u, v\} \cup S$ and define $\Psi = \Sigma_{I,I}$ and $\hat{\Psi} = \hat{\Sigma}_{I,I}$, two principal submatrices of size at most q . For the choice of ϵ from (4.6), the assumptions of Lemma 1 hold and we deduce that $\hat{\Psi}$ is invertible, with

$$(4.7) \quad \|\hat{\Psi}^{-1} - \Psi^{-1}\| < \frac{q\epsilon/\lambda^2}{1 - q\epsilon/\lambda} = \frac{qc}{(4+c)q + \lambda cq - \lambda cq} = \frac{c}{4+c}.$$

By Lemma 2, all diagonal entries of Ψ^{-1} are equal to one or greater, and so we can apply Lemma 3 with (4.7). Letting $\delta = c/(4+c)$, we get that

$$|\hat{\rho}_{uv|S} - \rho_{uv|S}| = \left| \frac{\hat{\Psi}_{uv}^{-1}}{\sqrt{\hat{\Psi}_{uu}^{-1}\hat{\Psi}_{vv}^{-1}}} - \frac{\Psi_{uv}^{-1}}{\sqrt{\Psi_{uu}^{-1}\Psi_{vv}^{-1}}} \right| < \frac{2\delta}{1-\delta} = \frac{c}{2}.$$

Therefore, $\|\hat{\Sigma} - \Sigma\|_\infty < \epsilon$ implies that our tests decide all conditional independences correctly in the RPC algorithm.

Next, using (2.8) and a union bound, we find that

$$\begin{aligned} \mathbb{P}\left(\hat{C}_\gamma(G) \neq C(G)\right) &\leq \mathbb{P}\left(|\hat{\Sigma}_{uv} - \Sigma_{uv}| \geq \epsilon \text{ for some } u, v \in V\right) \\ &\leq A \frac{p(p-1)}{2} \exp(-Bn\epsilon^2). \end{aligned}$$

Plugging in the definition of ϵ gives the claimed inequality

$$\mathbb{P}\left(\hat{C}_\gamma(G) \neq C(G)\right) \leq \frac{A}{2} p^2 \exp\left(-\frac{B\lambda^4 nc^2}{((4+c)q + \lambda cq)^2}\right) \leq \frac{A}{2} p^2 \exp\left(-\frac{B\lambda^4 nc^2}{36q^2}\right)$$

because $c \leq 1$ and $\lambda \leq 1$. The inequality $c \leq 1$ holds trivially because partial correlations are in $[-1, 1]$. The inequality $\lambda \leq 1$ holds because a $q \times q$ correlation matrix has trace q , this trace is equal to the sum of the q eigenvalues, and λ is the minimal eigenvalue. \square

5. SIMULATIONS

In this section we evaluate the finite-sample performance of the RPC algorithm in simulations. We compare RPC to two other versions of the PC-algorithm: (i) ‘Pearson-PC’, by which we mean the standard approach of using sample partial correlations to test Gaussian conditional independences, and (ii) ‘ Q_n -PC’, which is based on a robust estimator of the covariance matrix and was considered in [KB08]. All our computations are done with the `pcalg` package for R [KMC⁺12].

The Gaussian conditional independence tests in the `pcalg` package (and other software such as `Tetrad IV`¹) use a level $\alpha \in [0, 1]$ and decide that

$$(5.1) \quad X_u \perp\!\!\!\perp X_v | X_S \iff \sqrt{n - |S| - 3} \left| \frac{1}{2} \log \left(\frac{1 + \hat{\rho}_{uv|S}}{1 - \hat{\rho}_{uv|S}} \right) \right| \leq \Phi^{-1}(1 - \alpha/2).$$

If the observations are multivariate normal and $\hat{\rho}_{uv|S}$ are sample partial correlations then α is an asymptotic significance level for the test. The test in (5.1) is equivalent to our earlier setup of conditional independence tests in (3.1), with the exception of the sample size adjustment from n to $n - |S| - 3$. This adjustment is motivated by classical large-sample bias-correction theory for Fisher’s z-transform of sample correlations; compare [And03]. We show in the appendix that the adjustment has no affect on the consistency result we proved in Corollary 1.

Following the setup of [KB07], we simulate random DAGs and sample from probability distributions faithful to them. Fix a sparsity parameter $s \in [0, 1]$ and enumerate the vertices as $V = \{1, \dots, p\}$. Then we generate a DAG by including the edge $u \rightarrow v$ with probability s , independently for each pair (u, v) with $1 \leq u < v \leq p$. In this scheme, each node has the same expected degree, namely, $(p-1)s$.

Given a DAG $G = (V, E)$, let $\Lambda = (\lambda_{uv})$ be a $p \times p$ matrix with $\lambda_{uv} = 0$ if $u \rightarrow v \notin E$. Furthermore, let $\epsilon = (\epsilon_1, \dots, \epsilon_p)$ be a vector of independent random variables. Then the random vector X solving the equation system

$$(5.2) \quad X = \Lambda X + \epsilon$$

is well-known to be Markov with respect to G . Here, we draw the edge coefficients λ_{uv} , $u \rightarrow v \in E$, independently from a uniform distribution on the interval $(0, 1)$. For such a random choice, with probability one, the vector X solving (5.2) is faithful with respect to G . We consider three different types of data:

¹<http://www.phil.cmu.edu/projects/tetrad>

- (i) multivariate normal observations, which we generate by taking ϵ in (5.2) to have independent standard normal entries;
- (ii) observations from the Gaussian copula model, for which we transform the marginals of the normal random vectors from (i) to an $F_{1,1}$ -distribution;
- (iii) contaminated data, for which we generate the entries of ϵ in (5.2) as independent draws from a 80-20 mixture between a standard normal and a standard Cauchy distribution.

The contaminated distributions in (iii) do not belong to the nonparanormal class.

We consider the RPC algorithm in the version that uses Spearman correlations as in (2.6); the results for Kendall's τ are similar. When comparing graph estimates, we use the Structural Hamming Distance (SHD) as a measure of distance. The SHD is the number of edge insertions, deletions, and reorientations required to transform one graph to another. An undirected edge is counted as a single edge.

For the simulations we consider each combination of

$$p \in \{10, 22, 46, 100\} \quad \text{and} \quad n \in \{32, 100, 316, 1000, 3162\},$$

and choose the expected degree as either 3 or 6. In each case, we draw 240 random graphs and then generate samples of n observations. For the tuning parameter in (5.1), we consider a fixed grid, namely,

$$\log_{10} \alpha \in \{-7, -6, -5, -4.25, -3.5, -2.75, -2, -1.5, -1, -0.75\}.$$

For each of the resulting combinations, we run each of the three considered versions of the PC algorithm, retaining the result for the best choice among 7 values for α , best in terms of lowest average SHD to the true underlying DAG for a given combination. In Figures 1, 2 and 3, we plot the resulting SHDs against the sample size n .

A clear message emerges from the plots. First, Figure 1 shows that for normal data, RPC performs only marginally worse than Pearson-PC. The Q_n -PC algorithm also does well, although some gap in SHD arises for small sample sizes. Second, Figure 2 shows a dramatic gain in performance for RPC for the Gaussian copula data with $F_{1,1}$ marginals. In fact, the SHD associated with the other two graph estimators is comparable to that of estimating the graph to always be empty. The expected SHD between the empty graph and a graph on p nodes with expected degree d is simply the expected number of edges in our random graphs, which is $pd/2$. For our choices of $d = 3$ and $d = 6$, the respective expected SHD is 150 and 300 when $p = 100$. Finally, Figure 3 shows that RPC outperforms Q_n -PC for the contaminated data; Q_n -PC outperforms Pearson-PC for larger choices of p .

6. CONCLUSION

The PC algorithm of [SGS00] addresses the problem of model selection in graphical modelling with directed graphs via a clever scheme of testing conditional independences. For multivariate normal observations, the algorithm is known to have high-dimensional consistency properties when conditional independence is tested using sample partial correlations [KB07]. We show that the PC algorithm retains these consistency properties when observations follow a Gaussian copula model and rank-based measures of correlation are used to assess conditional independence. The assumptions needed in our analysis are no stronger than those in prior Gaussian work when the considered sequence of DAGs has bounded degree. When the degree grows our assumptions are slightly more restrictive as our proof requires control of

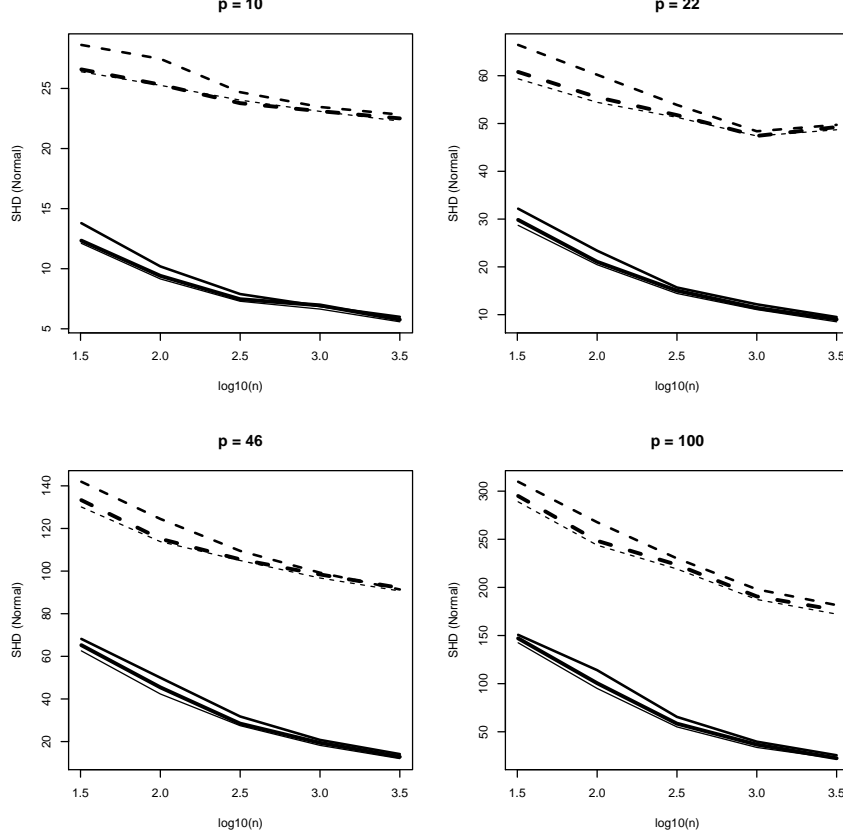


FIGURE 1. Structural Hamming distances for normal data, graphs with expected degree 3 (solid lines) and 6 (dotted lines), and three versions of the PC algorithm: Pearson-PC (thin lines), Q_n -PC (medium lines) and RPC using Spearman's ρ (thick lines).

the conditioning of principal submatrices of correlation matrices that are inverted to estimate partial correlations in the rank-based PC (RPC) algorithm.

Our simulations show that for normal data the RPC algorithm does essentially as well as the sample correlation-based version of the algorithm. As can be expected, we see RPC retain this performance for Gaussian copula data, for which sample correlations are poorly suited. Somewhat surprisingly, RPC also performed better than a previously considered robust version of the PC algorithm under a contamination model. We remark that the consistency theory available for this robust version is for a fixed graph size p . Since rank correlations take only marginally longer to compute than sample correlations, hardly any downsides are associated with making RPC the standard version of the PC algorithm for continuous data.

In our work on consistency, the data-generating distribution is assumed to be faithful to an underlying DAG. In fact, our results make the stronger assumption that non-zero partial correlations are sufficiently far from zero. As shown in

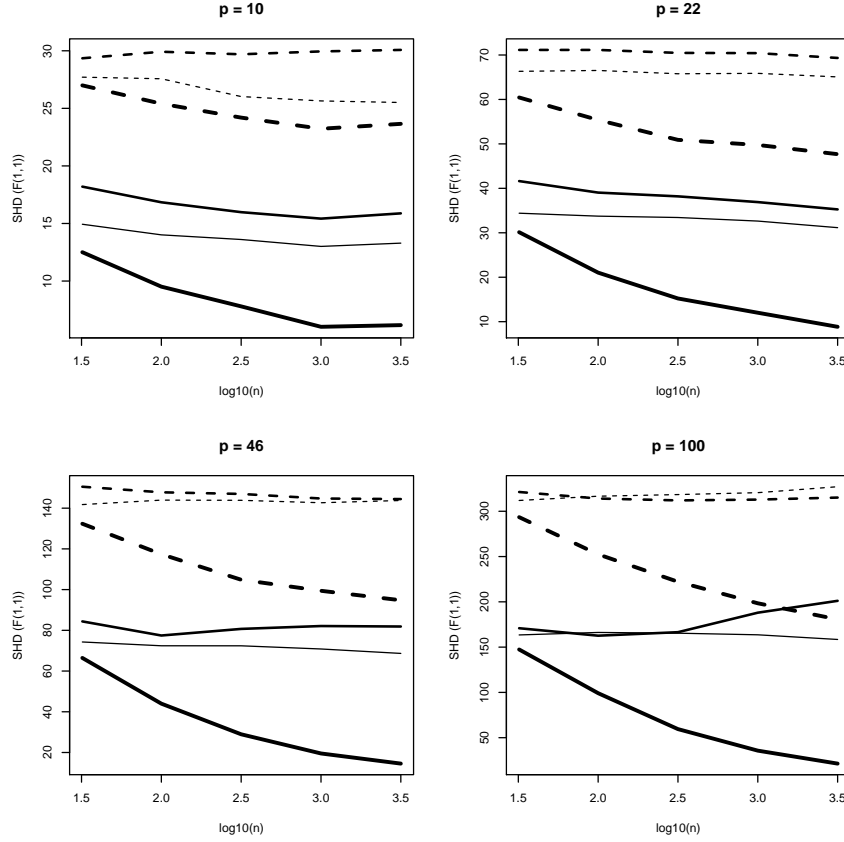


FIGURE 2. Structural Hamming distances for Gaussian copula data with $F_{1,1}$ marginals, graphs with expected degree 3 (solid lines) and 6 (dotted lines), and three versions of the PC algorithm: Pearson-PC (thin lines), Q_n -PC (medium lines) and RPC using Spearman's ρ (thick lines).

[URYB], this can be a restrictive assumption, which provides an explanation for why consistency does not ‘kick-in’ quicker in our simulation study.

Finally, we remark that extensions of the PC algorithm exist to deal with situations in which some causally relevant variables remain unobserved. Such algorithms infer a more complex graphical object; compare [SGS00] and [CMKR12]. It is reasonable to expect the use of rank correlations to be beneficial in those settings as well, and a study of these algorithms would be an interesting topic for future work.

ACKNOWLEDGMENTS

Mathias Drton was supported by the NSF under Grant No. DMS-0746265 and by an Alfred P. Sloan Fellowship.

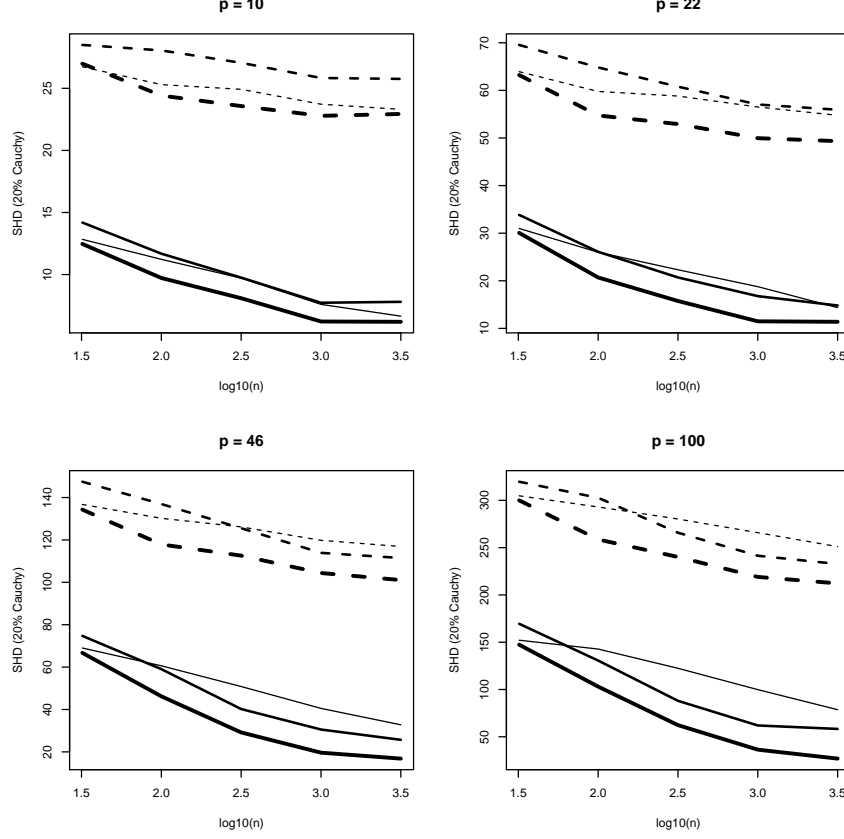


FIGURE 3. Structural Hamming distances for contaminated data, graphs with expected degree 3 (solid lines) and 6 (dotted lines), and three versions of the PC algorithm: Pearson-PC (thin lines), Q_n -PC (medium lines) and RPC using Spearman's ρ (thick lines).

APPENDIX A. SAMPLE SIZE ADJUSTMENT

We now show that the consistency result in Corollary 1 still holds when using the conditional independence tests from (5.1). In these tests, the sample size is adjusted from n to $n - |S| - 3$.

Proof. The test in (5.1) accepts a conditional independence hypothesis if and only if

$$(A.1) \quad |\hat{\rho}_{uv|S}| \leq \gamma(n, |S|, z),$$

where

$$(A.2) \quad \gamma(n, |S|, z) = \frac{\exp(z/\sqrt{n - |S| - 3}) - 1}{\exp(z/\sqrt{n - |S| - 3}) + 1}$$

and $z = z(\alpha) = 2\Phi^{-1}(1 - \alpha/2)$. We need to find a sequence (α_n) of values for α such that consistency holds under the scaling assumptions made in Corollary 1. We will do this by specifying a sequence (z_n) for values for the (doubled) quantiles z .

We claim that the RPC algorithm using the tests from (A.1) is consistent when choosing the quantile sequence

$$(A.3) \quad z_n = \sqrt{n-3} \cdot \log \left(\frac{1 + c_n/3}{1 - c_n/3} \right),$$

where we use the abbreviation

$$c_n := c_{\min}(\Sigma_n, q_n).$$

We will show that as the sample size n tends to infinity, with probability tending to one, $|\hat{\rho}_{uv|S} - \rho_{uv|S}| < c_n/3$ for every $u, v \in V$ and $|S| \leq q_n$. Furthermore, we will show that for the above choice of z_n and all sufficiently large n , we have $c_n/3 \leq \gamma(n, |S|, z_n) \leq 2c_n/3$ for each relevant set S with $0 \leq |S| \leq q_n$. These two facts imply that, with asymptotic probability one, every conditional independence test is correct, and the RPC algorithm succeeds.

First, we slightly adapt the proof of Theorem 2. Choosing the uniform error threshold for the correlation estimates as

$$(A.4) \quad \epsilon = \frac{c\lambda^2}{(6+c)q + \lambda cq} > 0$$

in place of (4.6) yields that, with probability at least

$$(A.5) \quad 1 - \frac{A}{2} p^2 \exp \left(-\frac{B\lambda^4 n c^2}{64q^2} \right),$$

we have that $|\hat{\rho}_{uv|S} - \rho_{uv|S}| < c/3$ for every $u, v \in V$ and $|S| \leq q$. When substituting p_n, q_n, c_n and $\lambda_{\min}(\Sigma_n, q_n)$ for p, q, c and λ , respectively, the scaling assumptions in Corollary 1 imply that the probability bound in (A.5) tends to one as $n \rightarrow \infty$, and we obtain the first part of our claim.

For the second part of our claim, note that our choice of z_n in (A.3) gives $\gamma(n, 0, z_n) = c_n/3$. Since $\gamma(n, |S|, z)$ is monotonically increasing in $|S|$, we need only show that for sufficiently large n ,

$$\gamma(n, q_n, z_n) - \gamma(n, 0, z_n) \leq c_n/3.$$

For $x \geq 0$, the function

$$f(x) = \frac{\exp(x) - 1}{\exp(x) + 1}$$

is concave and, thus, for any $q_n \geq 0$,

$$(A.6) \quad \begin{aligned} \gamma(n, q_n, z_n) - \gamma(n, 0, z_n) &= f \left(\frac{z}{\sqrt{n - q_n - 3}} \right) - f \left(\frac{z}{\sqrt{n - 3}} \right) \\ &\leq f' \left(\frac{z}{\sqrt{n - 3}} \right) \left(\frac{z}{\sqrt{n - q_n - 3}} - \frac{z}{\sqrt{n - 3}} \right). \end{aligned}$$

The derivative of f is

$$f'(x) = \frac{2\exp(x)}{(\exp(x) + 1)^2}.$$

Evaluating the right hand side of (A.6), we obtain that

$$\begin{aligned}
 \gamma(n, q_n, z_n) - \gamma(n, 0, z_n) &\leq \frac{1}{2} \left(1 - \frac{c_n^2}{9}\right) \log \left(\frac{1 + c_n/3}{1 - c_n/3}\right) \left(\frac{\sqrt{n-3}}{\sqrt{n-q_n-3}} - 1\right) \\
 (A.7) \qquad \qquad \qquad &\leq \frac{1}{2} \log \left(\frac{1 + c_n/3}{1 - c_n/3}\right) \left(\frac{\sqrt{n-3}}{\sqrt{n-q_n-3}} - 1\right).
 \end{aligned}$$

Being derived from absolute values of partial correlations, the sequence c_n is in $[0, 1]$. Now, $\log[(1+x)/(1-x)]$ is a convex function of $x \geq 0$ that is zero at $x = 0$ and equal to $\log(2)$ for $x = 1/3$. Therefore,

$$\frac{1}{2} \log \left(\frac{1 + c_n/3}{1 - c_n/3}\right) \leq \frac{1}{2} \log(2) \cdot c_n, \quad c_n \in [0, 1].$$

This shows that the bound in (A.7) is $o(c_n)$ because, by assumption, $q_n = o(\sqrt{n})$. In particular, the bound in (A.7) is less than $c_n/3$ for sufficiently large n , proving the claimed consistency result. \square

REFERENCES

- [AMP97] Steen A. Andersson, David Madigan, and Michael D. Perlman, *A characterization of Markov equivalence classes for acyclic digraphs*, Ann. Statist. **25** (1997), no. 2, 505–541.
- [And03] T. W. Anderson, *An introduction to multivariate statistical analysis*, third ed., Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2003.
- [Chi02] David Maxwell Chickering, *Learning equivalence classes of Bayesian-network structures*, J. Mach. Learn. Res. **2** (2002), no. 3, 445–498.
- [Chr05] David Christensen, *Fast algorithms for the calculation of Kendall’s τ* , Comput. Statist. **20** (2005), no. 1, 51–62.
- [CMKR12] Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson, *Learning high-dimensional directed acyclic graphs with latent and selection variables*, Ann. Statist. (2012), no. 40, 294–321.
- [DSS09] Mathias Drton, Bernd Sturmfels, and Seth Sullivant, *Lectures on algebraic statistics*, Oberwolfach Seminars, vol. 39, Birkhäuser Verlag, Basel, 2009.
- [HJ90] Roger A. Horn and Charles R. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1990, Corrected reprint of the 1985 original.
- [KB07] Markus Kalisch and Peter Bühlmann, *Estimating high-dimensional directed acyclic graphs with the PC-algorithm*, J. Mach. Learn. Res. **8** (2007), 613–636.
- [KB08] Markus Kalisch and Peter Bühlmann, *Robustification of the PC-algorithm for directed acyclic graphs*, J. Comput. Graph. Statist. **17** (2008), no. 4, 773–789.
- [KMC⁺12] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann, *Causal inference using graphical models with the R package pcalg*, Journal of Statistical Software **47** (2012), no. 11, 1–26.
- [Lau96] Steffen L. Lauritzen, *Graphical models*, Oxford Statistical Science Series, vol. 17, The Clarendon Press Oxford University Press, New York, 1996, Oxford Science Publications.
- [LHY⁺12] Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman, *High Dimensional Semiparametric Gaussian Copula Graphical Models*, [arXiv:1202.2169](#), 2012.
- [LLW09] Han Liu, John Lafferty, and Larry Wasserman, *The nonparanormal: semiparametric estimation of high dimensional undirected graphs*, J. Mach. Learn. Res. **10** (2009), 2295–2328.
- [Pea09] Judea Pearl, *Causality*, second ed., Cambridge University Press, Cambridge, 2009, Models, reasoning, and inference.
- [SGS00] Peter Spirtes, Clark Glymour, and Richard Scheines, *Causation, prediction, and search*, second ed., Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 2000, With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson, A Bradford Book.

- [URYB] Caroline Uhler, Garvesh Raskutti, Bin Yu, and Peter Bühlmann, *Geometry of faithfulness assumption in causal inference*, Manuscript.
- [VP91] Thomas Verma and Judea Pearl, *Equivalence and synthesis of causal models*, Tech. Report R-150, UCLA, 1991.

DEPARTMENT OF STATISTICS, THE UNIVERSITY OF CHICAGO, CHICAGO, IL, U.S.A.
E-mail address: `naftali@uchicago.edu`

DEPARTMENT OF STATISTICS, THE UNIVERSITY OF CHICAGO, CHICAGO, IL, U.S.A.
E-mail address: `drton@uchicago.edu`